DOCUMENT RESUME

ED 268 141 TM 850 786

AUTHOR	Kingston, Neal; And Others
TITLE	An Exploratory Study of the Applicability of Item Response Theory Methods to the Graduate Management Admission Test.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Graduate Management Admission Council, Princeton, NJ.
REPORT NO	ETS-RR-85-34
PUB DATE	Aug 85
NOTE	65p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	College Entrance Examinations; Computer Software; Correlation; *Equated Scores; Goodness of Fit; Graduate Students; *Graduate Study; Higher Education; *Latent Trait Theory; Mathematical Models; Mathematics Tests; Regression (Statistics); Scaling; *Scores; *Testing Programs; Test Items; Verbal Tests
IDENTIFIERS	*Graduate Management Admission Test; Item Parameters; *Linear Equating Method; Three Parameter Model

ABSTRACT

A necessary prerequisite to the operational use of item response theory (IRT) in any testing program is the investigation of the feasibility of such an approach. This report presents the results of such research for the Graduate Management Admission Test (GMAT). Despite the fact that GMAT data appear to violate a basic assumption of the three-parameter logistic item response model, local independence, the model was able to replicate accurately the observed item responses. IRT-based equating was consistent across two randomly selected samples and four selected subpopulations (male, female, younger examinees, and older examinees) and produced converted scores very similar to those produced by the current GMAT equating method--linear section pre-equating--a method that makes different assumptions than those required by IRT. It appears that for GMAT item types and populations, any effect of the violation of local independence on IRT true-score equating is negligible. This research has shown IRT equating to be feasible for the GMAT; but, because the local independence assumption of IRT appears to be violated, further experience is needed before other IRT methods--such as optimal test development using item information, or computerized adaptive testing--could be used for GMAT. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-85-34

ED268141

RESEARCH REPORT

AN EXPLORATORY STUDY OF THE APPLICABILITY OF ITEM RESPONSE THEORY METHODS TO THE GRADUATE MANAGEMENT ADMISSION TEST

Neal Kingston
Linda Leary
Larry Wightman

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

P. Feldmesser

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Educational Testing Service
Princeton, New Jersey
August 1985

An Exploratory Study of the Applicability of Item Response Theory Methods to the Graduate Management Admission Test[1]

Neal Kingston
Linda Leary
Larry Wightman

Educational Testing Service

This Research was Sponsored by the
Graduate Management Admission Council

August 1985

[1]The consultation and review of Daniel Eignor, Ronald Hambleton, Lawrence Hecht, Frederic Lord, Nancy Petersen, Martha Stocking, and Marilyn Wingersky is gratefully acknowledged. Special thanks to Louann Benton, Robin Durso, and Aster Tessema for their assistance in carrying out the data analyses. The opinions expressed herein are those of the authors and do not necessarily reflect those of Educational Testing Service, the Graduate Management Admission Council, nor any of the reviewers and consultants.

Copyright © 1985 by Graduate Management Admission Council and
Educational Testing Service. All rights reserved.

ABSTRACT

A necessary prerequisite to the operational use of item response theory (IRT) in any testing program is the investigation of the feasibility of such an approach. This report presents the results of such research for the Graduate Management Admission Test (GMAT).

Despite the fact that GMAT data appear to violate a basic assumption of the three-parameter logistic item response model, local independence, the model was able to replicate accurately the observed item responses. IRT-based equating was consistent across two randomly selected samples and four selected subpopulations (male, female, younger examinees, and older examinees) and produced converted scores very similar to those produced by the current GMAT equating method — linear section pre-equating — a method that makes different assumptions than those required by IRT. It appears that for GMAT item types and populations, any effect of the violation of local independence on IRT true-score equating is negligible.

This research has shown IRT equating to be feasible for the GMAT; but, because the local independence assumption of IRT appears to be violated, further experience is needed before other IRT methods — such as optimal test development using item information, or computerized adaptive testing — could be used for the GMAT.

EXECUTIVE SUMMARY

A necessary prerequisite to the operational use of item response theory (IRT) in any testing program is an investigation of the appropriateness of such an approach. The purpose of this research was to carry out such a feasibility study for the Graduate Management Admission Test (GMAT).

Two major approaches are used to assess the feasibility of using IRT methods for the GMAT. The first is to assess how well the assumptions of the model are fit by the data. But, the assumptions of a tractable psychometric model, such as IRT, will never be met exactly, because the human mind, which we are trying to model, is extremely complex.

Another approach is to gauge the extent to which violations of assumptions may preclude the use of a model. This can be done by investigating how well IRT can enhance particular important features of the GMAT. This approach is an inductive one, and, until a very large body of knowledge is developed, the possibility remains that one might find a set of conditions or tasks for which certain IRT-based methods will not work. Since these two approaches complement each other it is advisable to do both.

One attribute of major importance to the GMAT is the stability of the score scale as achieved through equating. Although the GMAT administers different editions of the test at different administrations, it is desirable that no examinee be advantaged by his or her choice of administration date. No matter how much effort one spends constructing tests trying to ensure that two editions of the test measure precisely the same attributes and are at precisely the same difficulty level, given sufficient data are available, it is nearly always possible to increase the comparability of scores with a statistical adjustment known as equating. GMAT equatings are based on data from one group of examinees, and the results of analyses of these data are applied to the scores of different groups of examinees. Thus, it is important that equatings be consistent across different naturally-occurring GMAT subpopulations.

This study addresses specifically two major concerns. How well are GMAT data (examinee population and item types) fit by the three-parameter logistic item response model? How consistent are IRT equatings across different non-random samples from the GMAT candidate population?

To answer these two questions, data were collected from two editions of the GMAT, to be referred to in this report as E3 and F1. E3 and F1 were administered preoperationally during the October 1981 administration. Six samples of GMAT examinees were drawn for the purpose of IRT equating edition F1 to edition E3. Two samples were random, one consisted of males exclusively, one of females exclusively, one of younger examinees (between the ages of 21 and 23) and the last of older examinees (29 years of age or older).

For each of the six samples, between 2,100 and 2,600 examinee responses to each item were analyzed for the three Verbal item types (reading comprehension, analysis of situations, and sentence correction) and two Quantitative item types (problem solving and data sufficiency) for each of the two test editions. IRT parameter estimates were obtained using LOGIST, a computer program designed expressly for this purpose. Parameter estimates for all six groups were placed on a common scale using a method referenced in the body of the report.

Assessing item response theory model fit is more an art than a science. As such, six different methods of assessing the reasonableness of the local independence and the three-parameter logistic form assumptions were used in this study: analysis of previous exploratory factor analysis, examination of item-type intercorrelations corrected for unreliability, analysis of item-ability regressions, analysis of modified Yen Q_1 statistics, comparison of item parameters estimated from heterogeneous and homogeneous subsets of items, and comparison of item parameters estimated for non-randomly selected subpopulations. The rationale and methods for these analyses are described in the body of this report. These analyses show that despite the fact that both the Verbal and Quantitative measures are multidimensional, examinee item responses are accurately modeled by a three-parameter logistic item response function.

Regarding the findings for the equatings, comparison of the converted scores for the IRT true-score equatings performed in the two random and four non-random groups shows that the equatings were very consistent for both the Verbal and Quantitative measures of the GMAT. In fact, at no raw score did the six converted scores vary more than one scaled score point. Since GMAT operational rounding practices frequently lead to reported score differences of this magnitude, the differences among the converted scores for the equatings performed on the six subgroups appear to be negligible. Also, empirically estimated standard errors of equating for IRT true score equating seem to be of about the same magnitude as the theoretical standard errors of equating for other equating procedures using similar data collection designs. Thus, an IRT true-score equating performed at one GMAT administration would be appropriate for the same test edition given at another administration, despite the likelihood that the IRT assumption of local independence is violated. In addition, the results of IRT true-score equating were very similar to those obtained operationally using SPE.

In summary, IRT true-score equating appears to produce consistent and accurate results, and thus appears to be appropriate for the Verbal and Quantitative measures of the GMAT. Perhaps other IRT methods will also be appropriate.

This research has provided evidence of the applicability of IRT-based methods to the GMAT, with both lines of inquiry pursued in the study yielding positive results. First, the IRT model seems to adequately fit GMAT data, despite the fact that the GMAT is designed as a heterogeneous test, an apparent violation of a key IRT assumption. Second, the various IRT-based equatings are consistent with one another and with the results of section pre-

equating. The positive nature of these findings has two major implications for the future of the GMAT, one relatively short-term, the other relatively long-term.

The major short-term implication of the research is that the GMAT program now has ample psychometric evidence to permit the development of operational IRT equating procedures. Such procedures, once developed, could be carried out concurrently with section pre-equating procedures to yield a second set of equating results. (Having a backup equating procedure is advisable for any testing program, so that if changes in test administration conditions become desirable, then the testing program is more likely to be in a position to respond.) Considering some of the potential advantages of IRT based equatings (see Appendix B for some of these advantages), these methods might actually prove superior to SPE methods in the long run.

The recommendations section of the report presents several other areas of research and development that are likely to present opportunities for improving the quality and reducing the costs of the GMAT program. The major long-term implication of the research is that the psychometric foundation has been set for the further exploration of a computer-adaptive (CAT) version of the GMAT. As IRT is the most powerful model for computer-adaptive testing, the fit of the IRT model to GMAT data demonstrated in this research provides the necessary first step in moving the GMAT toward the innovative assessment opportunities offered by CAT.

TABLE OF CONTENTS

	Page
Abstract	1
Executive Summary	ii
Introduction	1
Research Design	3
Description of the Test	3
Data Collection	4
Assessment of Model Fit	6
Analysis of Previous Exploratory Factor Analysis	7
Examination of Item Type Intercorrelations Corrected for Unreliability	9
Analysis of Item-Ability Regressions	11
Analysis of Modified Q_1 Statistics	14
Comparison of Item Parameters Estimated for Heterogeneous and Homogeneous Subsets of Items	16
Comparison of Item Parameters Estimated for Selected Subpopulations	18
Summary of Assessment of Model Fit	24
Comparison of Equatings	25
Rationale for Analyses	25
Comparison of Equatings for R1 and R2	27
Comparison of M,F,Y, and O Equatings with R1 Equating	30
Comparison of IRT Based Conversions with SPE Conversions	32
Summary of Equating Results	33
Recommendations	35
References	37
Appendix A - The Three-Parameter Logistic Item Response Model	39
Appendix B - Potential Advantages of Using IRT	41
Appendix C - Description of Samples	43
Appendix D - Parameter Estimates for E3 and F1 Items Calibrated in the R1 Group	45

INTRODUCTION

Before psychometric methods based item response theory (IRT) can be used operationally by any testing program, the feasibility of this approach needs to be investigated.[1,2] The purpose of this research was to carry out such a feasibility study for the Graduate Management Admission Test (GMAT).

Two major approaches are used to assess the feasibility of using IRT methods for the GMAT. The first is to assess how well the assumptions of the model are fit by the data. But, the assumptions of a tractable psychometric model, such as IRT, will never be met exactly, because the human mind, which we are trying to model, is extremely complex.

Another approach is to gauge the extent to which violations of assumptions may preclude the use of a model. This can be done by investigating how well IRT can enhance particular important features of the GMAT. This approach is an inductive one, and, until a very large body of knowledge is developed, the possibility remains that one might find a set of conditions or tasks for which certain IRT-based methods will not work. Since these two approaches complement each other it is advisable to do both.

One attribute of major importance to the GMAT is the stability of the score scale as achieved through equating. Although the GMAT administers different editions of the test at different administrations, it is desirable

[1] A description of item response theory and a discussion of its assumptions is given in Appendix A.

[2] The use of item response theory, if appropriate, can provide a testing program with numerous advantages. Some of these potential advantages are described in Appendix B.

that no examinee be advantaged by his or her choice of administration date. No matter how much effort one spends constructing tests trying to ensure that two editions of the test measure precisely the same attributes and are at precisely the same difficulty level, given sufficient data are available, it is nearly always possible to increase the comparability of scores with a statistical adjustment known as equating. GMAT equatings are based on data from one group of examinees, and the results of analyses of these data are applied to the scores of different groups of examinees who take the test edition at a later date. Thus, it is important that equatings be consistent across different naturally-occurring GMAT subpopulations.

This study addresses specifically two major concerns. How well are GMAT data (examinee population and item types) fit by the three-parameter logistic item response model? How consistent are IRT equatings across different non-random samples from the GMAT candidate population?

RESEARCH DESIGN

Description of the Test

The data presented in this report were obtained from the Graduate Management Admission Test, which reports Verbal, Quantitative, and Total scaled scores. The GMAT currently uses linear conversion parameters estimated by section pre-equating (Holland & Wightman, 1982) to place new editions of the test on scale. The reported scores are derived from six separately timed sections. The Verbal score is derived from reading comprehension, sentence correction and analysis of situations sections. The Quantitative score is derived from the problem solving (two sections) and data sufficiency sections. The Total score is based on all six sections and is equated separately. In order to implement section pre-equating, each test edition currently consists of eight sections: six operational sections that count toward an examinee's score (three verbal and three quantitative), and two preoperational sections that do not count toward an examinee's score (either both verbal, both quantitative or one verbal and one quantitative). Thus all of the examinees take all of the sections of the operational test, but only random[1] subsets of the examinees take each of the pairs of sections of the preoperational test. The preoperational test consists of six sections designed to be parallel in difficulty and content to their respective operational sections. An example

---

[1]Different versions of a test edition are packaged in an alternating fashion (e.g., 1,2,3,4, 1,2,3,4,...) referred to as spiralling. Research has shown that spiralling, when done correctly, results in essentially equivalent groups, sometimes even more effectively (due to a stratification effect) than does true random assignment.

of the complete structure of the predisclosure test edition used in this study is presented in Table 1. From this table it can be seen that Sections 2, 3, 4, 5, 7, and 8 were operational and Sections 1 and 6 were preoperational (or variable).

Table 1

Description of One Edition of the GMAT

Section	Content	# Items	Time
Operational Sections[1]			
1	(Variable)	—	30
2	Analysis of Situations (1)	20	20
3	Problem Solving	30	40
4	Reading Comprehension	25	30
5	Analysis of Situations (2)	20	20
6	(Variable)	—	30
7	Usage	25	15
8	Data Sufficiency	30	30
Preoperational Sections (placed in variable positions)			
-	Reading Comprehension	25	30
-	Analysis of Situations	35	30
-	Sentence Correction	25	30
-	Problem Solving (1)	20	30
-	Problem Solving (2)	20	30
-	Data Sufficiency	25	30

[1]Predisclosure format. Operational forms are now three each of 30 minute verbal and quantitative sections.

Data Collection

Two editions of the GMAT, 3EBS3 (hereupon referred to as E3) and 3FBS1 (referred to as F1) were administered preoperationally during the October 1981 administration of edition K-3BBS3 (B3). Six samples of GMAT examinees were drawn for the purpose of IRT equating edition F1 to edition E3.

Two samples were random (R1 and R2), one consisted of males exclusively (M), one of females exclusively (F), one of younger examinees, between the ages of 21 and 23 (Y), and the last of older examinees, 29 years of age or older (O).

Sixteen different versions of edition B3 contained a selection of preoperational test sections from edition E3 and 16 versions contained test sections from edition F1. Each item type appeared in five different E3 versions and in five different F1 versions. Every item type was paired with every other item type. Thus, some versions contained two verbal item-type preoperational sections, some two quantitative item-type preoperational sections, and some one verbal and one quantitative preoperational section. In order to obtain age group samples and sex group samples that are large enough for IRT equating, it was necessary to use data from all five appearances of the item type. It was decided to use the same sampling procedure for the two random samples so that the results from the six equatings would be comparable.

For each of the six samples, between 2,100 and 2,600 examinee responses to each item type in each test edition were analyzed. IRT parameter estimates were obtained separately for the verbal and quantitative items using LOGIST (Wingersky, Barton, & Lord, 1982; Wingersky, 1983). Parameter estimates for all verbal items were placed on a common metric using the TBLT method (Stocking & Lord, 1983). Likewise, the quantitative item parameter estimates were placed on a common metric. For each of the six samples, the exact sample sizes and the means and standard deviations of the raw scores for each of the six sections of the two editions of the test are presented in Appendix C.

ASSESSMENT OF MODEL FIT

Although there have been many attempts to develop a statistical test of model fit for the three-parameter logistic model, these attempts have not yet been successful. Most factor analytic techniques assume a linear relationship between items and factors. Item response theory allows for a nonlinear relationship. Estimation procedures for newer nonlinear factor analysis techniques are not yet widely available. Pearsonian chi-square methods, such as those used by Wright (1977) or Yen (1981), require expected (theoretical) frequencies based on true parameter values, but only estimates of those parameters are available. Likelihood ratio chi-square tests which, in general, have asymptotic properties, have been shown not to work with unconditional maximum likelihood estimation for the three-parameter model, even with reasonably large sample sizes (Lord, 1975). It should be noted, however, that the work of Rosenbaum (1984), which appeared after the analyses for this research were completed, appears very promising, with regard to a test of local independence, as does the work of Bock and Mislevy (personal communication with Frederic Lord).

Still, assessing model fit remains more an art than a science. As such, six different methods of assessing the reasonableness of the local independence[1] and the three-parameter logistic form assumptions were used in this study: analysis of previous exploratory factor analysis, examination of item-type intercorrelations corrected for unreliability, analysis of item-ability regressions, analysis of modified Q_1 statistics, comparison of item

[1]The assumption of local independence is met if the dimensionality of the IRT model is the same as the dimensionality of the data. See Lord and Novick (1968, chapter 16.3) for a fuller explanation of local independence.

parameters estimated from heterogeneous and homogeneous subsets of items, and comparison of item parameters estimated for non-randomly selected subpopulations.

Analysis of Previous Exploratory Factor Analysis

The three-parameter logistic item response model assumes unidimensionality, but it does not require the dimension to be linearly related to the variables (items) from which the latent trait is drawn. Although the theoretical underpinnings of a nonlinear factor analytic method were established almost 20 years ago, (McDonald, 1967), available technology allows us only to assess dimensionality through factor analysis with a linear model. Thus, although linear factor analytic results shed light on the fit of the model, they can not provide a definitive answer.

Swinton and Powers (1981) performed a principal axis factoring of the inter-item tetrachoric correlation matrices of three forms of the GMAT administered in November 1977. (It should be noted that since that time the makeup of the GMAT has been modified somewhat; these changes will be discussed later in this section.) An orthogonal varimax rotation was used in each of the three analyses. They retained six factors in each analysis. Five of the factors had the same (or very similar) interpretations across the three analyses. One factor was found in two of the three analyses and another was found in one analysis. It is likely that had more than six factors been retained, these additional factors would have appeared (as minor factors) in all analyses.

For the purpose of this report, a re-analysis of some of their results is presented. Table 2 shows, for each of the six item types, the percent of items (averaged across the three forms) with factor loadings above .30. From

this table it can be seen that the four verbal item types (reading comprehension, practical judgement - data evaluation, practical judgement - data applications, and English usage) load mainly on a dominant first verbal factor. Usage items load on a second strong verbal factor more strongly than they do on the first verbal factor. In addition, there are two relatively minor verbal factors. The two quantitative item types (problem solving and data sufficiency) each have a high proportion of items with loadings over .30 on the two quantitative factors. A seventh factor cuts across both the verbal and quantitative item types, appearing in practical judgement item types, problem solving, and data sufficiency.

Table 2

Summary of the Swinton and Powers GMAT Factor Analyses

Average Percent of Each Item Type with a Factor Loading of .30 or Greater on Each of the Seven Derived Factors

	Item Types					
Factors	RC	PJ1	PJ2	US	PS	DS
Verbal reasoning, comprehension	72	39	52	31	3	12
English usage	20	1	0	77	0	0
Practical Judgement answer key factor	0	34	0	3	0	0
Reading speed	30	4	2	14	0	0
Quantitative reasoning (algebra, geometry)	4	1	0	1	61	46
Quantitative reasoning (arithmetic)	0	0	0	0	44	32
Unnamed factor (maybe analytical reasoning)	0	16	63	0	24	17

RC - Reading Comprehension
PJ1 - Practical Judgement
PJ2 - Practical Judgement
US - Usage
PS - Problem Solving
DS - Data Sufficiency

Thus, it appears there are two major verbal factors, two major quantitative factors, and one unnamed factor that appears within both the Verbal and Quantitative measures, possibly analytical reasoning, in the GMAT as it was constituted in 1977. Since that time, the English usage items have been replaced with sentence correction items. It is not known whether or not this has altered the magnitude of the second verbal dimension. It is unlikely that this change would have created a unidimensional (in a linear sense) verbal scale. The second major change has been an increase in the proportion of problem solving items that constitute the Quantitative scale. Problem solving items that previously constituted 50 percent of the Quantitative items, now constitute 62 percent. Because both quantitative factors appear to be major in each of the Quantitative item types, the data do not suggest that this would have had any major influence on the factor structure.

Examination of Item Type Intercorrelations Corrected for Unreliability

The intercorrelations, corrected for unreliability, among the six sections constituting the Verbal and Quantitative scales, provide another type of evidence regarding the dimensionality of the GMAT. Table 3 presents intercorrelations and reliability estimates for editions E3 and F1, based on data from the operational administration of those editions. The lower triangle contains the intercorrelations, the diagonal contains the KR-20 reliability estimates, and the upper triangle contains the intercorrelations corrected for unreliability. The intercorrelations for the two quantitative item types are high, indicating that they are measuring very similar attributes. Note that the two problem solving sections are constructed from the same pool of items, and thus it is not surprising that they correlate highly with each other, and that each correlates the same with data sufficiency.

intercorrelations among the three verbal sections are considerably lower, particularly the correlations of analysis of situations (AS) with the other verbal sections, indicating the likelihood that these sections are measuring somewhat different attributes.

Table 3

Section Intercorrelations and Reliability Estimates[1]
GMAT Editions E3 and F1 Verbal and Quantitative Scales

E3

	Verbal			Quantitative		
	AS	RC	SC	DS	PS1	PS2
AS	.77	.66	.61	.59	.42	.40
RC	.51	.76	.82	.50	.31	.34
SC	.46	.62	.74	.53	.40	.40
DS	.44	.37	.39	.73	.90	.91
PS1	.32	.23	.30	.66	.72	.98
PS2	.29	.25	.29	.66	.70	.71

F1

	Verbal			Quantitative		
	AS	RC	SC	DS	PS1	PS2
AS	.82	.67	.70	.65	.55	.46
RC	.53	.75	.82	.51	.49	.38
SC	.54	.61	.74	.58	.55	.46
DS	.51	.38	.43	.74	.92	.92
PS1	.43	.37	.41	.69	.75	.97
PS2	.37	.30	.35	.70	.74	.79

[1]Intercorrelations appear in the lower triangle of each matrix, KR-20 reliability estimates on the diagonal, and intercorrelations corrected for unreliability in the upper triangle.

Analysis of Item-Ability Regressions

The analysis of item-ability regressions (IAR) is an exploratory graphical technique that compares the regression of the observed proportion of people getting an item correct on estimated theta (empirical regression) with the item response function based on the estimated item parameters (estimated regression).

The untransformed ability scale (theta estimated on the metric for which the trimmed calibration sample — examinees with estimated theta between -3.0 and 3.0 — has a mean of zero and a standard deviation of one) is split into 15 intervals of width .4 in the range -3.0 to +3.0. P_i, the proportion of people in interval i getting the item correct, adjusted for omits, is computed for each interval. That is,

(1) $$P_i = \frac{n_i^+ + n_i^o/A}{n_i}$$, where

n_i^+ is the number of examinees in the i-th interval who got the item correct,

n_i^o is the number of examinees in the i-th interval who omitted the item,

A is the number of alternatives per item, and

n_i is the number of examinees in interval i who answered the item or any item subsequent to that item.

The 15 P_i are plotted as squares for which areas are proportional to n_i. For each interval, a line of length $4\sqrt{PQ/n_i}$ is plotted, where P and Q are computed from the estimated item response function. The line is centered on the estimated response function. Although this line is a rough estimate of the .95 confidence interval around the item response functions, it is not being used as a statistical test. The reasons that this line does not represent the .95 confidence interval include: the use of the inappropriate

symmetric normal approximation to the binomial confidence interval around the response function (particularly a problem for extreme values of P); the use of an interval based on estimated item parameters; and the use of a line that is about 2% longer than it should be for a .95 confidence interval.

Figures 1A through 1D show four examples of item-ability regressions. The vertical scale in each is the probability of a correct response and ranges from 0 to 1. The horizontal scale is the ability metric and ranges from -3.0 to +3.0. Based on previous research (Kingston & Dorans, 1982), a model fit score was used to summarize each plot for the 170 verbal items and 130 quantitative items calibrated in the randomly chosen group of examinees, R1.

Insert Figure 1 About Here

For each of the 15 groups of thetas, the number of times that the midpoint of the box representing the empirical proportion correct does <u>not</u> intersect the vertical line representing the .95 confidence interval is counted. Thus Figure 1A has a model fit score of 0, Figure 1B has a model fit score of 1, Figure 1C has a model fit score of 2 and Figure 1D has a model fit score of 3, the highest model fit score obtained in this study. Table 4 presents the cumulative proportion of model fit scores broken down by the five GMAT item types, as well as, in parentheses, the number of items with each model fit score.

Table 4

Analysis of Item-Ability Regressions

Item Type	Number of Items	Model Fit Score Cumulative Proportion 0	1	2	3
All Verbal	170	.67(114)	.90(39)	1.00(11)	1.00(6)
Reading Comprehension	50	.70(35)	.94(12)	.98(2)	1.00(1)
Analysis of Situations	70	.61(43)	.81(14)	.93(8)	1.00(5)
Sentence Correction	50	.72(36)	.98(13)	1.00(1)	
All Quantitative	130	.69(90)	.93(31)	.98(7)	1.00(2)
Problem Solving	80	.76(61)	.95(15)	.99(3)	1.00(1)
Data Sufficiency	50	.58(29)	.90(16)	.98(4)	1.00(1)

The IAR model fit score has been shown to be insensitive to many types of multidimensionality (Kingston & Dorans, 1982), so high model fit scores are likely only to indicate lack of logistic form for the item response function.

If the vertical lines in the item-ability regressions did represent conditional .95 confidence intervals, if the assumptions of the three-parameter logistic model were met, and if we had true parameter values instead of estimates, then the expected model fit score would be above .75 (that is, 15 times .05). Since this is not the case, a normative approach has been used, with scores of 2 or above considered indicative of a possible lack of model fit.

Comparisons with previous research using verbal and quantitative item types from the Graduate Record Examinations Aptitude Test (renamed the GRE General Test in 1981), show that overall, the model fit scores for the GMAT verbal item types are slightly better than those for GRE verbal item types and populations, while scores for the GMAT quantitative item types are considerably better than those for GRE quantitative item types and populations. Perhaps this is due to a greater homogeneity of the GMAT population. In any

case, this analysis indicates that estimated three-parameter logistic item response functions can replicate GMAT data very well, even though the GMAT measures appear to be multidimensional.

Examination of Table 4 shows that there are fewer low model fit scores for the analysis of situations item type than for the other two verbal item types. The statistical significance of this difference was assessed using a chi-square test of independence. Model fit scores of two or more were combined into a single grouping because when expected cell frequencies are too small, the test statistic is not distributed chi-square. The chi-square was significant at about the .04 level. A similar analysis for the two quantitative item types was not conclusive, with the chi-square significant at only the .09 level.

Analysis of Modified Q_1 Statistics

Yen (1981) developed a statistic, referred to as Q_1, to assess the fit of the three-parameter logistic model to test data. Further research (Yen, 1984) showed that the statistic was distributed approximately chi-square, and that Q_1 was not sensitive to violations of local independence.

The modified Q_1 statistic is described in equation 2.

$$(2) \quad Q_{1i} = \sum_{j=1}^{17} [N_j(O_{ij} - E_{ij})^2]/[E_{ij}(1 - E_{ij})], \text{ where:}$$

N_j is the number of examinees in cell j,

O_{ij} is the observed proportion of examinees in cell j that passes item i, and

E_{ij} is the predicted proportion of examinees in cell j that passes item i. E_{ij} is the mean P_i in cell j based on the estimated theta of the examinees in cell j and the estimated parameters of the item.

This statistic differs from Yen's in that it uses 17 groups based on equal intervals along the theta metric with equal intervals of width .4, except for the first group, theta less than -3, and the last group, theta greater than +3. Yen used 10 groups chosen so as to have approximately equal sample sizes. With the modified Q_1, if the number of examinees in a cell is less than or equal to five, that cell is collapsed into the adjoining cell. The (approximate) degrees of freedom are the number of cells minus the number of parameters estimated. Since the degrees of freedom ranged from 15 (no cells collapsed, and c-parameters not estimated) to 12 (two cells collapsed, and all three parameters estimated), the probabilities of a Type I error associated with the value of Q $[P(Q_1)]$, although not strictly correct, were examined rather than the values of the test statistic itself. Table 5 presents, for each item type, the distribution of $P(Q_1)$, broken down into four categories: .00-.05, .06-.25, .26-.50, and .51-1.00. A low value of $P(Q_1)$ indicates relatively poor model fit. The proportion of items in each category is indicated in parentheses.

Table 5

Distribution of P's Associated with Q_1

	$P(Q_1)$			
	.00-.05	.06-.25	.26-.50	.51-1.00
All Verbal	21(.12)	32(.19)	27(.16)	90(.53)
Reading Comprehension	6(.12)	8(.16)	12(.24)	24(.48)
Analysis of Situations	13(.19)	14(.20)	5(.07)	38(.54)
Sentence Correction	2(.04)	10(.20)	10(.20)	28(.56)
All Quantitative	13(.10)	21(.16)	19(.15)	77(.59)
Problem Solving	5(.06)	8(.10)	14(.18)	53(.66)
Data Sufficiency	8(.16)	13(.26)	5(.10)	24(.48)

Table 5 shows a higher proportion of low $P(Q_i)$ than would be expected if the assumptions of the model were met and the test statistic was distributed chi-square. This is so for three of the five item types: reading comprehension, analysis of situations, and data sufficiency. The findings for the analysis of situations and data sufficiency item types are consistent with those from the analysis of item-ability regressions.

Comparison of Item Parameters Estimated for Heterogeneous and Homogeneous Subsets of Items

If a test is multidimensional, then parameter estimates based on a subset of items that is factorially more homogeneous than the total measure will differ from parameter estimates based on the factorially heterogeneous total test (Bejar, 1980; Kingston & Dorans, 1982). For unidimensional tests, the local independence assumption will be met, so the parameters will be invariant. Using the data for the R1 group, for each item type, the relationship between parameter estimates based on the total verbal measure were compared with those based on just that item type. Table 6 presents, for each item type, the correlation[1] between the item parameters estimated for heterogeneous and homogeneous subsets of items, and the mean and standard deviation of the parameter estimates.

Table 6 shows that only the a-parameter estimates consistently behaved differently for the heterogeneous and homogeneous calibrations. For each item type, the mean a is greater for the homogeneous calibration. This finding is

[1]Correlating b-parameter estimates is not strictly appropriate and can be misleading because extreme b-parameters have large standard errors of estimate. Since these extreme values also have the greatest weight in the correlation formula (by squaring deviates their influence is increased) this can lower the correlation even if the relationship between the actual parameters is perfect. If the correlation is high, however, a strong case is made for the consistency of the b-parameters.

consistent with Reckase's hypothesis (1979) that in the face of multidimensional data, LOGIST parameter estimates are for a theta metric which is the centroid of the separate factors. Similar results were found by Kingston and Dorans (1982) in their analysis of the GRE verbal measure.

Table 6

Comparison of Item Parameters Estimated from Heterogeneous and Homogeneous Subsets of Items

Parameter		Item Types RC	AS	SC	PS	DS
A	correlation	.80	.82	.91	.94	.98
	mean, heterogeneous	.58	.63	.62	.75	.78
	mean, homogeneous	.71	.72	.72	.78	.86
	s.d., heterogeneous	.22	.29	.24	.28	.28
	s.d., homogeneous	.24	.35	.26	.27	.34
B	correlation	.99	.91	.96	.98	1.00-
	mean, heterogeneous	-.42	.04	-.05	-.07	.12
	mean, homogeneous	-.38	.13	-.17	-.07	.04
	s.d., heterogeneous	1.41	1.23	1.25	1.26	1.60
	s.d., homogeneous	1.30	1.48	1.14	1.15	1.50
C	correlation	.69	.76	.69	.76	.96
	mean, heterogeneous	.15	.19	.16	.15	.14
	mean, homogeneous	.19	.20	.15	.16	.15
	s.d., heterogeneous	.08	.12	.11	.11	.09
	s.d., homogeneous	.09	.13	.11	.10	.09

Note that, although the correlation between the b-estimates for the analysis of situations items is relatively low, only .91, this is due to a single item. This item was difficult (the b was about 3 for the heterogeneous calibration, and was about 8 for the homogeneous calibration) and had a very large standard error of estimate. With this one item removed, the correlation between the b-estimates would have been .96.

When comparing parameter estimates based on heterogeneous and homogeneous subsets of items, two things should be remembered. First, this technique will only find evidence of multidimensionality when the factors are

primarily in only one of the item subsets. That is, if the additional factor cuts across each subset, then it will not affect parameter estimates. In this analysis, homogeneous subsets of items were based on item type. Alternative subsets based on content might have produced different results.

Related to this is a second consideration. Even if the comparison of heterogeneous and homogeneous parameter estimates reveals multidimensionality, if test forms are developed to be fairly parallel, then the effect of multidimensionality on parameter estimates will be consistent and IRT-based methods, particularly test equating, might still be be appropriate.

Comparison of Item Parameters Estimated for Selected Subpopulations

In order to assess the possible effect of multidimensionality on the use of IRT for a testing program that develops fairly parallel test editions, parameters could be estimated for selected subpopulations that differ with respect to various characteristics. This was done as part of this study (see the data collection section of this report for more details), and parameters were estimated for two random groups (R1 and R2, used for comparison purposes) and four selected groups: males (M), females (F), younger (Y, ages 21-23), and older (O, ages 29+).

Table 7 presents the correlations among the estimated b-parameters for each item type for the six samples. With two sets of exceptions, the correlations among the b-parameter estimates are all high and, for the most part, there is little difference between the correlations among the non-random subgroups and the correlation between the two random groups. The first exception occurs for the correlations with the estimates made for problem solving items in the female sample. As stated earlier, extreme b-parameters have large standard errors, which can lead to a low correlation between estimated b's. This is what occurred here. Among the problem solving items there was one

item that was particularly difficult, with a b-parameter estimated as about 2 in five of the six groups. In the female group, however, parameter estimation difficulties were exacerbated by a relative dearth of quantitatively very able examinees, and the b-parameter was estimated as about 39. This lowered the correlation between b-parameters estimated in the female group and those estimated in the other groups to about .43. The F* line in the tables presents the correlations among estimated b-parameters with the one very poorly estimated parameter left out. Those correlations are in line with those in the rest of the table. It should be noted that the overestimation of the b-parameter was compensated for with an underestimation of the a-parameter. Thus the proportion correct estimated at each theta is not greatly affected, and the effect on equating of this inaccurate estimation is minimized. Similarly a single item, for which the b-parameter estimates ranged from about -1 to 18 in the six groups, accounted for the low correlations among the b-estimates for the reading comprehension items. With that item excluded, the correlations for reading comprehension would have been in line with these for the other item types.

Table 8 presents the correlations among the a-parameter estimates, by item type, derived in each of the six groups. The correlations are lower than those among the estimated b-parameters because a-parameters have a larger standard error of estimate than do b-parameters. In general, the correlations among the a-parameters estimated in the four non-random groups are not all that different from the correlation between the estimated a-parameters in the two random groups.

Table 9 presents the correlations among the c-parameter estimates for the six groups. As expected, since c-estimates have fairly large standard errors, these correlations are typically lower than those for the a- and

b-parameter estimates. Some of the correlations are extremely low (.12, .16, etc.). This is an artifact of the LOGIST estimation procedure. LOGIST uses a criterion value (called CRITCFIX), related to item difficulty, to decide whether or not there are enough data to estimate c for an item. If not, a common c is estimated for that item along with all other items for which there are not enough data. If an item has a criterion value at about the critical point, one time it may have its own c estimated and another time it may be assigned the common c value. If, as occurred here, a few items had large c values when estimated independently, but the common c had a relatively low value (as it typically does), then the correlations among estimated c-parameters can be severely depressed.

Table 7

Correlations Among Estimated b-Parameters
By Item Type for Six Groups

Problem Solving

	R1	R2	F	M	O	Y
R1	1.00	0.97	0.45	0.98	0.97	0.93
R2	0.97	1.00	0.43	0.97	0.96	0.95
F	0.45	0.43	1.00	0.43	0.43	0.41
F*	0.96	0.98	1.00	0.94	0.93	0.96
M	0.98	0.97	0.43	1.00	0.98	0.93
O	0.97	0.96	0.43	0.98	1.00	0.89
Y	0.93	0.95	0.41	0.93	0.89	1.00

Data Sufficiency

	R1	R2	F	M	O	Y
R1	1.00	0.98	0.97	0.99	0.94	0.98
R2	0.98	1.00	0.99	0.98	0.98	0.99
F	0.97	0.99	1.00	0.96	0.98	0.99
M	0.99	0.98	0.96	1.00	0.95	0.97
O	0.94	0.98	0.98	0.95	1.00	0.96
Y	0.98	0.99	0.99	0.97	0.96	1.00

Reading Comprehension

	R1	R2	F	M	O	Y
R1	1.00	0.73	0.95	0.87	0.93	0.68
R2	0.73	1.00	0.86	0.96	0.50	0.99
F	0.95	0.86	1.00	0.94	0.84	0.83
M	0.87	0.96	0.94	1.00	0.70	0.93
O	0.93	0.50	0.84	0.70	1.00	0.43
Y	0.68	0.99	0.83	0.93	0.43	1.00

Analysis of Situations

	R1	R2	F	M	O	Y
R1	1.00	0.95	0.91	0.93	0.92	0.95
R2	0.95	1.00	0.94	0.96	0.93	0.94
F	0.91	0.94	1.00	0.95	0.92	0.96
M	0.93	0.96	0.95	1.00	0.97	0.96
O	0.92	0.93	0.92	0.97	1.00	0.93
Y	0.95	0.94	0.96	0.96	0.93	1.00

Sentence Correction

	R1	R2	F	M	O	Y
R1	1.00	0.97	0.96	0.93	0.94	0.93
R2	0.97	1.00	0.98	0.96	0.96	0.94
F	0.96	0.98	1.00	0.94	0.95	0.93
M	0.93	0.96	0.94	1.00	0.96	0.97
O	0.94	0.96	0.95	0.96	1.00	0.95
Y	0.93	0.95	0.93	0.97	0.95	1.00

Table 8

Correlations Among Estimated a-Parameters
by Item Type for Six Groups

Problem Solving

	R1	R2	F	M	O	Y
R1	1.00	0.84	0.73	0.89	0.89	0.83
R2	0.84	1.00	0.67	0.92	0.91	0.86
F	0.73	0.67	1.00	0.65	0.65	0.71
M	0.89	0.92	0.65	1.00	0.92	0.85
O	0.89	0.91	0.65	0.92	1.00	0.80
Y	0.83	0.86	0.71	0.85	0.80	1.00

Data Sufficiency

	R1	R2	F	M	O	Y
R1	1.00	0.83	0.77	0.90	0.83	0.82
R2	0.83	1.00	0.89	0.89	0.87	0.92
F	0.77	0.89	1.00	0.79	0.84	0.93
M	0.90	0.89	0.79	1.00	0.88	0.84
O	0.83	0.87	0.84	0.88	1.00	0.86
Y	0.82	0.92	0.93	0.84	0.86	1.00

Reading Comprehension

	R1	R2	F	M	O	Y
R1	1.00	0.80	0.82	0.83	0.79	0.77
R2	0.80	1.00	0.87	0.90	0.84	0.89
F	0.82	0.87	1.00	0.84	0.83	0.87
M	0.83	0.90	0.84	1.00	0.92	0.84
O	0.79	0.84	0.83	0.92	1.00	0.78
Y	0.77	0.89	0.87	0.84	0.78	1.00

Analysis of Situations

	R1	R2	F	M	O	Y
R1	1.00	0.89	0.80	0.87	0.86	0.79
R2	0.89	1.00	0.78	0.88	0.86	0.82
F	0.80	0.79	1.00	0.76	0.75	0.84
M	0.87	0.88	0.76	1.00	0.94	0.82
O	0.86	0.86	0.75	0.94	1.00	0.76
Y	0.79	0.82	0.84	0.82	0.76	1.00

Sentence Correction

	R1	R2	F	M	O	Y
R1	1.00	0.89	0.91	0.74	0.86	0.79
R2	0.89	1.00	0.90	0.81	0.87	0.78
F	0.91	0.90	1.00	0.74	0.84	0.77
M	0.74	0.81	0.74	1.00	0.83	0.78
O	0.86	0.87	0.84	0.83	1.00	0.78
Y	0.79	0.78	0.77	0.78	0.78	1.00

Table 9

Correlations Among Estimated c-Parameters
by Item Type for Six Groups

Problem Solving

	R1	R2	F	M	O	Y
R1	1.00	.64	.64	.76	.80	.57
R2	.64	1.00	.76	.84	.79	.72
F	.64	.76	1.00	.58	.68	.72
M	.76	.84	.58	1.00	.86	.59
O	.80	.79	.68	.86	1.00	.61
Y	.57	.72	.72	.59	.61	1.00

Data Sufficiency

	R1	R2	F	M	O	Y
R1	1.00	.63	.68	.80	.77	.69
R2	.63	1.00	.79	.59	.83	.71
F	.68	.79	1.00	.48	.84	.86
M	.80	.59	.48	1.00	.55	.50
O	.77	.83	.84	.55	1.00	.85
Y	.69	.71	.86	.50	.85	1.00

Reading Comprehension

	R1	R2	F	M	O	Y
R1	1.00	.21	.39	.38	.60	.39
R2	.21	1.00	.36	.50	.31	.58
F	.39	.36	1.00	.57	.79	.70
M	.38	.50	.57	1.00	.76	.37
O	.60	.31	.79	.76	1.00	.49
Y	.39	.58	.70	.37	.49	1.00

Analysis of Situations

	R1	R2	F	M	O	Y
R1	1.00	.65	.47	.58	.66	.67
R2	.65	1.00	.59	.66	.72	.60
F	.47	.59	1.00	.63	.59	.72
M	.58	.66	.63	1.00	.81	.70
O	.66	.72	.59	.81	1.00	.65
Y	.67	.60	.72	.70	.65	1.00

Sentence Correction

	R1	R2	F	M	O	Y
R1	1.00	.59	.44	.23	.46	.29
R2	.59	1.00	.72	.46	.51	.31
F	.44	.72	1.00	.16	.29	.12
M	.23	.46	.16	1.00	.45	.50
O	.46	.51	.29	.45	1.00	.43
Y	.29	.31	.12	.50	.43	1.00

Summary of the Assessment of Model Fit

There are two assumptions made by the item response model used in this study: (1) each of the two GMAT measures, Verbal and Quantitative, are unidimensional, and (2) the regression of the probability of a correct response on theta has a logistic form that can be described using no more than three parameters. The analysis of the exploratory factor analysis and the comparison of item parameters estimated for homogeneous and heterogeneous subsets of items indicate that both the Verbal and the Quantitative measures almost surely are multidimensional. Each measure probably has two major dimensions, and possibly a number of minor ones.

The analysis of item-ability regressions and the comparison of item parameter estimates based on the diverse subpopulations indicate that the regression of the probability of a correct response on the estimated theta (which appears to be a composite of the underlying dimensions) is accurately modeled by a three-parameter logistic item response function. The analysis of the modified Q_1 statistics, however, indicates a possible minor departure from logistic form for reading comprehension, analysis of situations, and data sufficiency items. Since the actual distribution of Q_1 is not chi-square, this departure may be an artifact of the analysis.

COMPARISON OF EQUATINGS

The Verbal and Quantitative measures of the GMAT edition F1 were each equated to the respective measures in edition E3, using IRT true-score equating (Lord, 1980, chapter 13.5)[1]. These equatings were done for each of the six samples described earlier: random 1, random 2, males, females, younger and older.

Rationale for Analyses

If the assumptions of an IRT model are met, then IRT true-score equating is appropriate. If the logistic form assumption is met, but the local independence assumption is violated, IRT equating might still be appropriate—that is, it might still provide accurate equating.

The three-parameter logistic model, when applied to the GMAT, accurately reflected examinee responding behavior at the item level. This is supported by the analysis of item-ability regressions and the modified Q_1 statistic. Even if the assumption of local independence is violated, this indicates that IRT can accurately predict the mean score of a test and, thus, the overall relative difficulties of two tests for a given population. This occurs because the mean score of a test is equal to the sum of the mean item scores. The violation of local independence would mean, however, that an individual examinee's score is not simply the sum of the probabilities of a correct response for all of the items in the test. Thus, if local independence is violated, although IRT might allow one to estimate accurately the

[1]Total scores were not equated, since if IRT equating were used operationally, the Total scaled score would be a linear composite of the verbal and quantitative scaled scores:

$$\text{Total} = 5 * (V + Q) + 200.$$

mean score of a test, it would not allow an accurate estimation of the distribution of test scores for a group of examinees. For example, when local independence is violated, IRT equating might underestimate the variance of the score distribution. But, if different editions of a test are parallel with respect to the factors that underlie the test, then any inaccuracies in the estimation of the score distributions for one edition of the test might be compensated for by a corresponding inaccuracy in the estimated score distribution for the other edition of the test. Thus, the equating relationship determined by the use of IRT might still be correct.

The presence of differences in factor structure over different editions of the test and naturally-occuring subpopulations (that is, different test administration dates) and the effect that these differences might have on equating need to be assessed to decide if IRT equating is appropriate in the face of multidimensionality. This can be done by selecting diverse subpopulations on which to equate the test and comparing those equatings. The diversity of the subpopulations acts to magnify any differences in the factor structure of the two editions of the test that are being equated.

If the assumption of local independence is violated, could IRT true-score equatings be consistent across diverse populations, but consistently incorrect? Such an occurrence, although unlikely, is a possiblity. Unfortunately, there is no good criterion to judge the correctness of the IRT true-score equatings. For example, if the IRT equatings were shown to differ from the operational SPE equatings, either one of the equating methods could be "right" and the other one "wrong." If, on the other hand, the two equating methods did agree, then it is that much more likely that each of the methods is producing an essentially correct result, especially since the two methods make very different assumptions.

Comparison of Equatings for R1 and R2

The Verbal and Quantitative measures of the GMAT edition F1 were equated to those for E3 twice, once in each of two randomly selected groups of examinees who took the items from those tests at the October 1981 administration. The F1 scores were then placed on the GMAT scale using the linear scaling parameters determined for edition E3. Tables 10 and 11 present the equated F1 scores for selected raw scores for the Verbal and Quantitative measures, respectively. The tables also include the standard deviation of the equated scores across the two groups. This is an estimate of the empirical standard error of equating (S.E.E.). In the last column, for comparison purposes, is a theoretical S.E.E. This last S.E.E. is not based on the IRT true-score equating model used in this research, because a formula for such a standard error has not yet been developed. Instead, the S.E.E. is based on a linear equating model using a common item random-groups design (Angoff, 1971 Design III), as this is the data collection design most similar to the one used in this study.[1] It is expected that the theoretical standard error of the IRT true score equating based on this data collection design is actually somewhat larger than this, since more parameters have to be estimated than need be estimated for a linear equating model.

Table 10 shows that the difference between the R1 and R2 verbal equatings was smaller than the standard error of equating expected with a linear common item equating design, an equating design for which there is no compar-

[1]The calculation of the S.E.E. assumed a raw score mean and standard deviation for F1 of 40.1 and 15.3, respectively, and a scaled score standard deviation of 9.0. These values based on the Test Analysis for this form (Wightman, 1982). Also, the correlation between anchor test and the new edition was assumed to be .85.

able issue of violation of local independence. Table 11 shows that for the quantitative measure, except at the extremes of the score scale, the R1 and R2 equating are as consistent or more consistent than would be expected with the linear common item design. Although at the extremes there is more error, the theoretical S.E.E. reported is an underestimate of the S.E.E. of an IRT true score equating for a test that met the assumption of local independence. Thus, with random groups, the violation of local independence does not appear to affect the equating.

Operational GMAT equatings would be derived in one group but applied in one or more other similar but not randomly equivalent groups. Thus, the R1-R2 comparison is a lower bound to the kind of variability in an equating function that might occur in an operational GMAT equating scheme.

Table 10

Verbal Measure
Unrounded Converted Scores and Standard Errors of Equating
for Selected Raw Scores

							Standard Errors of Equating		
Raw Score [1]	Converted Scores [2] R1	R2	Male	Female	Younger	Older	Empirical Random [3]	Empirical Non Random [4]	Empirical Linear [5]
0	.65	.83	-1.05	-.89	-1.01	.55	.13	.77	.35
10	9.01	8.96	8.42	8.35	8.89	8.96	.04	.31	.28
20	16.47	16.47	16.70	16.46	16.85	16.77	.00	.17	.21
30	22.83	22.97	23.29	23.19	23.50	23.16	.10	.15	.16
40	28.57	28.75	28.85	29.04	29.16	28.66	.13	.22	.13
50	34.04	34.21	33.99	34.32	34.38	33.72	.12	.31	.16
60	39.59	39.79	39.32	39.51	39.74	38.90	.14	.36	.21
70	45.71	45.98	45.46	42.21	45.53	45.10	.19	.20	.28
80	52.14	52.21	51.91	51.44	51.54	52.29	.05	.39	.35

Table 11

Quantitative Measure
Unrounded Converted Scores and Standard Errors of Equating
for Selected Raw Scores

							Standard Errors of Equating		
Raw Score [1]	Converted Scores [2] R1	R2	Male	Female	Younger	Older	Empirical Random [3]	Empirical Non Random [4]	Empirical Linear [5]
0	8.15	9.09	8.60	8.64	9.14	9.05	.66	.28	.31
10	15.11	15.18	15.14	15.16	15.22	15.27	.05	.06	.23
20	21.46	21.38	21.55	21.25	21.24	21.41	.06	.15	.16
30	28.04	28.05	28.11	27.75	27.67	28.16	.01	.25	.12
40	35.00	35.18	35.07	34.90	34.84	35.36	.13	.23	.15
50	42.45	42.95	42.79	42.87	42.91	43.06	.35	.11	.21
60	50.53	51.32	51.18	51.34	51.42	51.65	.56	.20	.29

[1] Formula scored - Rights - Wrongs/4
[2] On GMAT 0-60 reported score scale
[3] Standard deviation (n-1) of R1 and R2 converted scores
[4] Standard deviation (n-1) of M, F, Y, and O converted scores
[5] Angoff (1971) equation 22, also see text

Comparison of M,F,Y, and O Equatings with R1 Equatings

Tables 10 and 11 also include the converted scores for the Verbal and Quantitative equatings, respectively, based on the male, female, younger, and older subpopulations, as well as an empirical estimate of the standard error of equating for those non-random groups. For the Verbal measure, other than at a raw score of 0, 50, and 60, the empirical standard error is very similar to the theoretical linear equating standard error, which is smaller than the correct but incalculable theoretical IRT standard errors of equating for a unidimensional test.

The differences among the converted scores should be viewed in terms of GMAT score reporting practices. GMAT uses formula scoring, R-W/4, rounding to the nearest integer (with unrounded scores ending in .5 always rounded in the examinee's favor). For the Verbal measure, this introduces differences in scaled scores between rounded and unrounded raw formula scores of up to about .3 scaled score points. In addition, equated scores are rounded to the nearest integer for score reporting. This introduces differences of up to .5 scaled score points. Thus, a total unrounded "wobble" of up to about .8 scaled score points exists with current GMAT score reporting practices. This means that for a given raw score, the reported score might be one scaled score point different than the unrounded, obtained equated score.

Table 10 shows that even when the effect of multidimensionality is magnified by selection of non-random subpopulations, the differences in reported scores would be no more than 1 point (for raw score of Ø, reported scores would be either 1 (for R1, R2, and O) or Ø (for M, F, and Y), since Ø is the lowest reported score).

Figures 2 and 3 present the results of the Verbal equatings graphically. Figure 2 shows the six conversion lines. They are so close to each other that they are, for the most part, indistinguishable. Figure 3 shows the differences between each conversion line and the R1 line by subtracting the R1 equated score from each other group's equated score at each raw score point. This serves to magnify any differences between the equatings. These figures confirm the consistency among the equatings found in Table 10.

Insert Figure 2 About Here

Insert Figure 3 About Here

Table 11 shows a consistency among the Quantitative equatings similar to that found for the Verbal equatings. The empirical standard errors of equating are, for the most part, small compared to the theoretical linear S.E.E., and the reported converted scores never differ by more than one scaled score point. In fact, at raw scores of 10, 30, and 40, all equatings yield the same reported score.

Figures 4 and 5 are similar to Figures 2 and 3. Figure 4 presents the Quantitative conversion lines, and Figure 5 presents the differences between each conversion line and the R1 conversion line. Again, all the equatings appear consistent.

Insert Figure 4 About Here

Insert Figure 5 About Here

Comparison of IRT Based Conversions with SPE Based Conversions

Figures 6 and 7 compare the IRT based conversions for edition F1 (based on the R1 group) with the section pre-equating based conversions for the GMAT Verbal and Quantitative measures, respectively. There are several caveats that apply to these comparisons. First, the section pre-equating method used for the GMAT allows only a linear conversion, while IRT equating as used in this study produces a curvilinear conversion. Second, the path of the equatings and, thus, the potential sources of error, are different for the two methods. For IRT, F1 was equated to E3 and then placed on the GMAT scale through the section pre-equating of E3 to two old editions, KB2 and KB3. For SPE, F1 was equated directly to KB2 and KB3. Third, the IRT equating and section pre-equating are based on data from different administrations (But, evidence regarding the consistency of IRT equatings across non-random subgroups indicates that the effect of this should be minor). In any case, if the results of IRT equating and section pre-equating are similar under these less-than-perfect conditions, then it is likely that they would be at least as similar under more reasonable circumstances.

Figure 6 shows that for most of the raw score range, the IRT equating and SPE differ by no more than one scaled score point. It should be noted

that where the difference between the two equatings is larger, below a raw score of about nine, there are very few examinees. When edition F1 was administered operationally in January 1983, only about two percent of the examinees had a raw score of nine or below.

Figure 7 shows that for the Quantitative measure, for most of the raw score range, the IRT equating and SPE are in very close agreement. For raw scores between 0 and 48, the corresponding scaled scores differ by less than one-half a point. Only between raw scores of 57 and 65 did scaled score differences between the equatings start to get large enough to be of any significance, two to four points. And, as stated earlier, there is no way of determining, from available data, which of the equatings is closer to the truth.

Insert Figure 6 About Here

Insert Figure 7 About Here

Summary of Equating Results

Comparison of the converted scores for the IRT true-score equatings performed in the two random and four non-random groups shows that the equatings were very consistent for both the Verbal and Quantitative measures of the GMAT. In fact, at no raw score did the six converted scores vary more than one scaled score point. Since GMAT operational rounding practices frequently lead to reported score differences of this magnitude, the differences among

the converted scores for the equatings performed on the six subgroups appear to be negligible. Also, the standard error of equating for IRT true-score equating seems to be of about the same magnitude as the theoretical standard error of equating for other equating procedures using similar data collection designs. Thus, we expect that an IRT true-score equating performed at one GMAT administration would be appropriate for the same test edition given at another administration, despite the likelihood that the IRT assumption of local independence is violated.

The results of IRT true-score equating were very similar to those obtained operationally using SPE. Thus it appears very unlikely that IRT equating of a multidimensional GMAT Verbal or Quantitative measure might be consistent but incorrect.

In summary, IRT true-score equating produces consistent and accurate results, and thus appears to be appropriate for the Verbal and Quantitative measures of the Graduate Management Admission Test.

RECOMMENDATIONS

This research has provided evidence of the applicability of IRT-based methods to the GMAT, with both lines of inquiry pursued in the study yielding positive results. First, the IRT model seems to adequately fit GMAT data, despite the fact that the GMAT is designed as a heterogeneous test, an apparent violation of a key IRT assumption. Second, the various IRT-based equatings are consistent with one another and with the results of section pre-equating. The positive nature of these findings leads to several recommendations regarding the development and integration of IRT methods into the GMAT program.

One immediate implication of this research is that the GMAT program now has ample psychometric evidence to permit the development of operational IRT equating procedures that parallel section pre-equating (SPE) and could be carried out concurrently to yield a second set of equating results. Assessing the consistency of results of these two different methods would allow a powerful quality assurance check of the current equating process. Development of operational IRT equating procedures would also position the GMAT program to change equating methods should legal or administrative constraints be imposed on the currently used data collection design. One part of this development would be the determination of a formula for a Total score that is a relatively simple composite of the IRT equated Verbal and Quantitative scores (Total score would not be equated separately using IRT, as it is with SPE, because of the gross multidimensionality of the Total measure). Total scores would be based on composites of Verbal and Quantitative scores weighted so as to ensure comparability with the existing SPE equated Total score.

Another recommended project is the assessment of IRT equating based on pretest data. Currently GMAT items are administered at least three times experimentally prior to their operational use. With IRT based equating, it might be possible to cut the number of pre-operational administrations to one or two. In addition, if results were favorable, an IRT equating system based on using pretest data would require that only one experimental section be administered to each examinee, rather than the current two. Also, operational sections could be based on items that come from a large variety of pretests, and thus no examinees would be exposed previously (due either to their repeatedly taking the GMAT or to a test form becoming non-secure) to more than a few items in the operational part of the test. Thus, an IRT pretest-data based equating system, if psychometrically feasible, might increase the security of test editions, reduce program costs, and reduce examination time.

IRT has potential major advantages in areas other than test equating. A third recommended project is the development of test specifications based on IRT parameters. This would allow the GMAT program to make use of the ETS IRT test development system (the system is scheduled to be operational in July 1985). When appropriate, use of IRT can make the test development process more efficient, leading to reduced costs, and can also result in more efficient measurement itself, by maximizing the information yielded by the test.

Two other recommended avenues of research and development have somewhat longer range potential for leading to cost reductions or program improvements: (1) use of simplified IRT scoring to improve the reliability of GMAT scores simultaneous with the reduction of the length of the test, and (2) computerized adaptive testing. GMAC efforts are already underway to explore the applicability of computerized adaptive testing (CAT) to the GMAT. The results of this research provide evidence of the psychometric soundness of this approach.

REFERENCES

Bejar, I.I. (1980) A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement, 17, 283-296.

Holland, P.W. & Wightman, L.E. (1982) Section pre-equating: a preliminary investigation. In Holland, P.W. & Rubin, D.B. (editors) Test Equating. New York: Academic Press.

Kingston, N.M. & Dorans, N.J. (1982) The feasibility of using item response theory as a psychometric model for the GRE aptitude test. ETS Research Report 82-12. Princeton, N.J.: Educational Testing Service.

Lord, F.M. (1975) Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters. ETS Research Bulletin 75-33. Princeton, N.J.: Educational Testing Service.

Lord, F.M. (1980) Application of item response theory to practical testing problems. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Lord, F.M. & Novick, M.R. (1968) Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

McDonald, R.P. (1967) Nonlinear factor analysis. Psychometric Monographs, No. 15.

Reckase, M.D. (1979) Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4, 207-230.

Rosenbaum, P.R. (1984) Testing the conditional independence and monotonicity assumptions of item response theory. Psychometrika, 49, 425-435.

Stocking, M.L. & Lord, F.M. (1983) Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Swinton, S.S. & Powers, D.E. (1981) Construct validity of the Graduate Management Admission Test: a factor analytic study. GMAC Research Report 81-1. Princeton, NJ: Educational Testing Service.

Wingersky, M.S. (1983) LOGIST: program for computing maximum likelihood procedures for logistic test models. In Hambleton, R.K. (editor) Applications of item response theory. Vancouver, B.C.: Educational Research Institute of British Columbia.

Wingersky, M.S.; Barton, M.A.; Lord, F.M. (1982) LOGIST Users Guide. Princeton, N.J.: Educational Testing Service.

Wightman, L.E. (1983) Test analysis: Graduate Management Admission Test Form 3FBS1. Statistical Report 83-84. Princeton, NJ: Educational Tesing Service.

Wright, B. (1977) Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14 97-116.

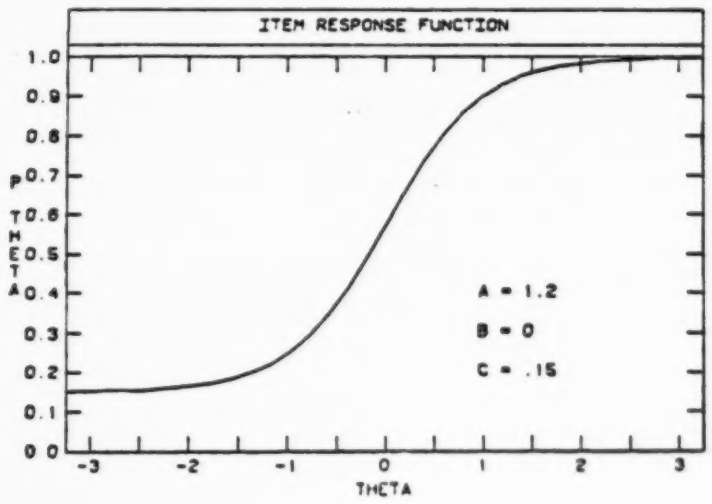
Yen, W.M. (1981) Using simulation results to choose a latent trait model. Applied Psychological Measurement. 5, 245-262.

Yen, W.M. (1984) Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8, 125-145.

APPENDIX A

The Three-Parameter Logistic Item Response Model

The three-parameter logistic item response model postulates that underlying examinees' responses to test items is a single unobservable (but not inestimable) ability. (Ability is used in a narrow psychometric sense throughout this paper. It refers to any latent trait which can be used to order examinees. It makes no assumption concerning when or how this trait was learned or developed.) The probability of an examinee with a particular level of ability (θ) responding correctly to an item depends solely on three parameters associated with the item, a (the ability of the item to differentiate among examinees of different ability levels), b (the difficulty of the item), and c (the probability of a very low ability examinee responding correctly). The following figure presents an item response function labeled with its parameters.



The a, b, and c parameters determine the relationship between ability and probability of a correct response according to the following equation:

$$P_g(\theta) = c_g + (1-c_g)/[1+e^{-1.7a_g(\theta-b_g)}].$$

There are two separate assumptions in this equation: there is a single ability (there is only one θ in the equation), and the relationship between ability and probability of a correct response has a logistic form requiring no more than three parameters (the form of this equation is referred to as logistic in the statistical community).

APPENDIX B

Potential Advantages of Using IRT

At least seven major benefits may accrue to the GMAT from the use of item response theory:

1. Possible improvement in maintenance of GMAT scales. All equatings attempt to approximate an ideal with varying degrees of success. IRT may provide a better equating than section pre-equating, thus yielding a more effective method of maintaining the stability and meaning of the GMAT scales across forms of the test.
2. Utility for innovative assessment. Item response theory is the most powerful available model for computerized adaptive testing (that is, using a computer to tailor the test to the individual examinee's ability level by selecting only the most appropriate items). Thus, if the GMAC decides to explore adaptive testing for the GMAT, a natural first step would be to investigate the feasibility of IRT methods for the GMAT.
3. Reduction of lead time required for changes in test content. During the life of many national programs, the need arises to introduce new item types or to substantially modify existing item types. Since IRT item calibrations can take place during the pretest data analysis phase, the use of IRT analyses could reduce the lead time required between the time an item is first written and when it is administered operationally.

4. Reduction in testing time. Use of IRT methods would require only one variable section, rather than the two which are currently used with section pre-equating. Thus, the testing time for GMAT could be reduced from four to three-and-one-half hours. Also, fewer subforms would be required, thereby simplifying the test production process.

5. Facilitation of test development. IRT provides a variety of methods that would allow the analysis of the statistical properties of a test form before it is administered operationally. These methods can be used during the assembly of a test form to ensure that predetermined statistical specifications are met (e.g., difficulty level, discrimination, and standard errors of measurement at particular score points).

6. Reduction in item exposure. Using IRT methods, test items might need to be administered only twice, once for pretesting and calibration and once in the operational form. In current practice, section pre-equating is used with three or four administrations of each item.

7. Flexibility in item placement by administration. Section pre-equating requires that all intact sections of a form being equated be administered pre-operationally during the same administration. For IRT equating methods, items subsequently assembled for a final operational form need not have been pretested (nor calibrated) at the same administration. Thus, test security would be improved considerably, since at no one administration would any or all of the test booklets contain all of the items that would comprise one operational final form.

APPENDIX C

Table C1
Description of Samples

		Edition E3						Edition E2					
		R1	R2	M	F	Y	O	R1	R2	M	F	Y	O
RC	$\bar{x}$	14.56	14.56	14.23	15.03	14.84	14.41	12.28	12.45	11.94	13.04	12.38	12.68
	s	5.00	5.05	5.04	4.79	4.61	5.25	5.29	5.23	5.28	5.11	4.91	5.57
	n	2,435	2,439	2,396	2,564	2,413	2,234	2552	2,510	2,527	2,465	2,387	2,213
AS	$\bar{x}$	16.87	16.97	16.42	17.97	18.31	15.68	16.04	16.27	15.49	17.21	17.55	14.98
	s	6.60	6.70	6.54	6.75	6.70	6.23	7.70	7.60	7.69	7.78	7.82	7.54
	n	2,465	2,439	2,510	2,451	2,387	2,218	2,506	2,560	2,482	2,505	2,377	2,186
SC	$\bar{x}$	11.97	11.91	11.51	12.86	12.42	11.91	11.21	11.30	10.69	12.38	11.70	11.46
	s	5.26	5.31	5.21	5.14	4.86	5.44	5.40	5.35	5.27	5.25	5.20	5.60
	n	2,584	2,486	2,518	2,440	2,460	2,138	2,491	2,455	2,464	2,498	2,401	2,170
PJ1	$\bar{x}$	8.46	8.61	9.19	7.53	8.87	8.37	9.29	9.03	9.70	8.13	9.70	8.74
	s	4.02	3.97	4.13	3.59	3.98	4.12	4.47	4.49	4.57	4.23	4.41	4.61
	n	2,491	2,518	2,478	2,551	2,369	2,139	2,496	2,462	2,521	2,504	2,417	2,176
PJ2	$\bar{x}$	8.36	8.49	9.16	7.02	8.78	8.08	2.11	9.02	9.63	7.82	9.34	8.79
	s	4.02	4.17	4.27	3.65	4.11	4.30	4.62	4.50	4.59	4.18	4.56	4.53
	n	2,460	2,584	2,501	2,491	2,463	2.127	2,459	2,475	2,497	2,497	2,478	2,224
DS	$\bar{x}$	9.77	9.83	10.42	8.87	10.36	9.32	10.78	10.83	11.44	9.99	11.43	10.26
	s	4.83	4.92	5.08	4.44	5.09	4.72	4.70	4.71	4.87	4.38	4.65	4.73
	n	2,521	2,547	2,473	2,501	2,392	2.160	2,471	2,379	2,428	2,494	2,425	2,137

Table C2

Difference in Standard Errors of the Mean
Between Random and Non Random Groups[1]

	Edition E3				Edition F1			
	M	F	Y	O	M	F	Y	O
RC	-4.6	6.5	3.9	- 2.1	-5.7	9.1	0.2	4.3
AS	-5.3	11.1	14.6	-13.1	-6.2	9.8	13.0	-10.9
SC	-5.8	12.4	6.5	- .4	-7.4	14.7	5.8	2.7
PJ1	11.6	-17.8	5.9	- 2.9	8.5	-16.2	8.5	- 6.6
PJ2	12.5	-23.9	6.0	- 5.9	8.7	-19.2	4.2	- 4.2
DS	9.1	-13.6	8.2	- 7.0	9.4	-12.1	9.3	- 8.1

[1]Non random group mean minus combined R1+R2 mean, divided by the standard error of the combined R1+R2 mean.

Appendix D

Parameter Estimates for E3 and F1 Items Calibrated in the R1 Group Analysis of Situations

Item #	Parameter A	Parameter B	Parameter C
1	0.64	0.44	0.50
2	0.75	-0.42	0.22
3	1.04	1.39	0.50
4	0.47	-1.17	0.11
5	0.61	0.28	0.19
6	0.32	-1.47	0.11
7	0.72	0.37	0.41
8	0.76	0.97	0.24
9	0.15	-1.11	0.11
10	0.28	-1.13	0.11
11	1.00	0.22	0.46
12	0.30	-1.07	0.11
13	0.93	0.03	0.26
14	0.83	-0.18	0.32
15	0.54	-1.61	0.11
16	0.40	-0.91	0.11
17	0.34	0.12	0.11
18	0.38	-1.20	0.11
19	0.73	0.94	0.32
20	0.42	1.22	0.10
21	0.51	-2.25	0.11
22	0.61	-1.30	0.11
23	0.82	1.23	0.19
24	0.46	-2.08	0.11
25	0.47	0.88	0.17
26	0.44	0.06	0.11
27	0.27	-2.41	0.11
28	0.67	3.47	0.18
29	1.10	1.43	0.20
30	0.59	-1.23	0.11
31	0.95	1.40	0.20
32	0.70	2.27	0.26
33	0.71	0.26	0.28
34	1.30	2.26	0.13
35	0.22	0.99	0.11
36	0.37	-1.73	0.11
37	0.49	0.64	0.11
38	0.57	-1.12	0.11
39	0.43	-0.91	0.11
40	1.05	0.33	0.23
41	0.34	-2.09	0.11
42	1.27	0.55	0.37
43	0.33	-0.19	0.11
44	1.21	1.74	0.08
45	0.40	0.23	0.11
46	0.58	-1.05	0.11
47	1.30	0.73	0.11
48	0.33	0.19	0.11
49	0.39	1.15	0.10
50	0.78	0.51	0.31
51	0.55	0.51	0.22
52	0.53	0.80	0.16
53	0.91	0.16	0.23
54	0.84	-0.12	0.11
55	0.10	1.29	0.11
56	0.69	0.07	0.50
57	0.44	-0.47	0.11
58	0.53	-0.45	0.11
59	0.69	1.61	0.33
60	0.24	-2.77	0.11
61	0.61	1.24	0.22
62	0.75	1.41	0.45
63	0.85	0.09	0.43
64	0.96	0.59	0.06
65	0.59	0.75	0.30
66	0.63	-0.78	0.11
67	1.16	0.85	0.10
68	0.43	-1.70	0.11
69	0.43	-0.23	0.11
70	0.79	0.24	0.39

Data Sufficiency

Item #	Parameter A	Parameter B	Parameter C
1	0.51	-3.36	0.07
2	0.94	-1.65	0.07
3	0.56	-1.93	0.07
4	0.64	-1.22	0.07
5	0.16	-3.81	0.07
6	0.69	-0.92	0.00
7	0.91	-0.18	0.27
8	0.36	-2.21	0.07
9	0.49	-1.51	0.07
10	1.13	0.61	0.19
11	0.78	0.23	0.26
12	0.72	0.35	0.14
13	0.60	0.25	0.07
14	1.23	0.45	0.29
15	0.37	1.04	0.07
16	0.47	0.73	0.00
17	0.63	1.49	0.11
18	0.63	0.86	0.07
19	1.39	1.69	0.24
20	1.17	1.35	0.08
21	0.67	1.77	0.14
22	0.86	2.27	0.15
23	0.75	1.70	0.14
24	1.07	1.61	0.12
25	1.01	2.14	0.08

Item #	Parameter A	Parameter B	Parameter C
26	0.72	-2.24	0.07
27	0.81	0.00	0.47
28	0.69	-1.84	0.07
29	0.75	-1.61	0.07
30	0.89	-0.64	0.33
31	0.58	-1.34	0.07
32	0.74	-0.51	0.20
33	0.82	-0.02	0.25
34	0.36	-2.22	0.07
35	0.96	0.07	0.05
36	0.36	-1.40	0.07
37	0.95	0.48	0.18
38	0.43	0.55	0.07
39	0.54	-1.87	0.07
40	0.51	0.15	0.07
41	1.04	1.35	0.27
42	0.73	1.06	0.19
43	1.34	1.37	0.25
44	0.74	1.20	0.14
45	1.23	2.89	0.17
46	1.23	1.86	0.29
47	0.80	1.92	0.15
48	1.05	1.72	0.15
49	0.77	1.43	0.10
50	1.22	2.04	0.08

Problem Solving

Item #	Parameter A	Parameter B	Parameter C	Item #	Parameter A	Parameter B	Parameter C
1	0.29	-2.95	0.07	41	0.75	-1.97	0.07
2	0.41	-2.93	0.07	42	0.73	-1.60	0.07
3	0.94	-1.21	0.33	43	0.64	-1.55	0.07
4	0.88	-0.24	0.18	44	0.81	-0.62	0.24
5	0.42	-1.15	0.07	45	0.39	0.04	0.07
6	0.74	0.60	0.36	46	0.85	-0.37	0.48
7	0.35	-0.35	0.07	47	0.87	-0.75	0.23
8	0.44	-0.30	0.07	48	0.78	-0.05	0.09
9	0.55	-1.06	0.07	48	0.45	-1.49	0.07
10	0.82	1.02	0.36	50	0.61	-0.53	0.00
11	0.52	-1.96	0.07	51	0.98	0.31	0.18
12	1.02	1.28	0.22	52	0.50	0.80	0.11
13	0.65	0.49	0.16	53	0.29	-1.00	0.07
14	0.82	0.61	0.07	54	0.70	1.05	0.24
15	1.04	2.11	0.37	55	1.02	0.64	0.09
16	0.95	0.81	0.09	56	1.16	1.11	0.16
17	1.01	0.81	0.19	57	0.48	2.12	0.35
18	0.98	1.67	0.28	58	0.65	1.19	0.04
19	0.92	0.42	0.09	59	1.01	0.91	0.21
20	0.65	1.68	0.02	60	0.53	0.87	0.00
21	0.56	-2.70	0.07	61	0.36	-2.63	0.07
22	0.56	-1.77	0.07	62	1.12	-1.21	0.43
23	0.29	-1.39	0.07	63	0.36	-2.70	0.07
24	0.75	-1.01	0.07	64	0.86	-0.57	0.19
25	0.93	-0.23	0.06	65	0.94	-0.03	0.36
26	0.35	-1.12	0.07	66	0.79	-1.41	0.07
27	0.31	-1.37	0.07	67	1.11	0.35	0.32
28	0.31	-0.86	0.07	68	0.56	-1.41	0.07
29	0.39	-1.17	0.07	69	1.04	0.48	0.27
30	1.05	0.10	0.18	70	0.59	-1.29	0.07
31	0.51	-0.09	0.00	71	1.01	0.22	0.14
32	0.55	1.26	0.20	72	0.56	0.40	0.05
33	0.73	0.61	0.30	73	0.94	0.25	0.06
34	0.88	0.95	0.12	74	1.32	0.57	0.10
35	0.55	1.09	0.08	75	1.09	1.11	0.11
36	1.40	1.64	0.36	76	0.83	1.54	0.19
37	1.35	0.82	0.26	77	1.20	1.28	0.17
38	0.75	1.28	0.17	78	1.12	1.35	0.28
39	0.92	1.13	0.02	79	0.88	0.93	0.18
40	0.73	1.18	0.16	80	0.49	-0.45	0.07

Reading Comprehension

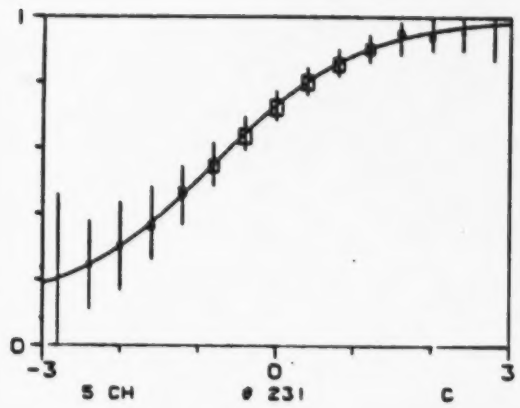
Item #	Parameter A	Parameter B	Parameter C
1	0.64	-2.15	0.11
2	1.00	-1.72	0.11
3	0.29	-0.65	0.11
4	0.90	-0.52	0.18
5	0.66	-1.06	0.11
6	0.47	-0.71	0.11
7	0.57	-0.92	0.11
8	0.75	-0.05	0.38
9	0.38	-0.21	0.11
10	0.20	-0.92	0.11
11	0.23	-1.30	0.11
12	0.52	-2.00	0.11
13	0.60	-2.35	0.11
14	0.45	-2.46	0.11
15	0.65	-0.04	0.13
16	0.43	-2.35	0.11
17	0.50	-1.07	0.11
18	0.86	-0.51	0.41
19	0.37	0.93	0.11
20	0.35	1.37	0.02
21	0.74	-2.51	0.11
22	0.62	1.53	0.35
23	0.87	1.85	0.23
24	0.51	1.42	0.06
25	0.58	-1.17	0.11
26	0.43	-1.17	0.11
27	1.06	-0.46	0.21
28	0.50	0.86	0.35
29	0.99	-0.67	0.08
30	0.45	1.67	0.17
31	0.66	-2.36	0.11
32	0.71	-0.76	0.11
33	0.44	-1.33	0.11
34	0.39	-1.55	0.11
35	0.41	2.45	0.18
36	0.50	-1.08	0.11
37	0.53	-0.59	0.11
38	0.58	-0.39	0.11
39	0.57	-1.03	0.11
40	0.58	2.39	0.35
41	0.06	1.78	0.11
42	0.57	-1.93	0.11
43	0.61	-0.99	0.11
44	0.70	-0.12	0.17
45	0.74	0.29	0.10
46	0.38	-1.72	0.11
47	0.45	-0.31	0.11
48	0.78	2.14	0.12
49	1.13	2.27	0.26
50	0.50	-0.77	0.11

Sentence Correction

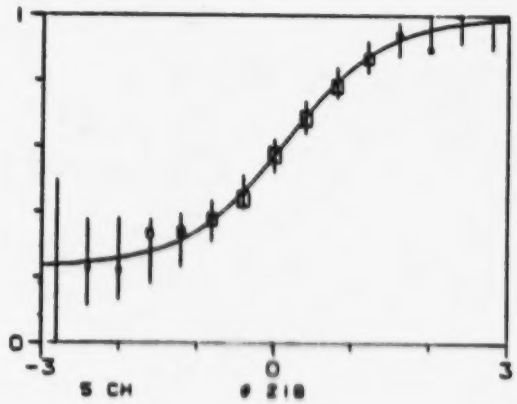
Item #	Parameter A	Parameter B	Parameter C	Item #	Parameter A	Parameter B	Parameter C
1	0.66	-2.27	0.11	26	0.46	-2.22	0.11
2	0.67	-1.72	0.11	27	0.27	-1.80	0.11
3	0.84	-1.67	0.11	28	0.45	-1.62	0.11
4	0.63	-1.52	0.11	29	0.73	-0.57	0.50
5	0.35	-2.44	0.11	30	0.76	-0.25	0.31
6	0.64	0.67	0.50	31	0.49	-0.76	0.11
7	0.65	-1.42	0.11	32	0.55	-1.16	0.11
8	0.33	-1.98	0.11	33	0.36	-1.18	0.11
9	0.34	-0.56	0.11	34	0.52	0.07	0.11
10	0.57	-0.51	0.11	35	0.55	-0.36	0.11
11	0.89	-0.29	0.11	36	1.00	-0.08	0.21
12	1.30	0.16	0.20	37	0.65	1.42	0.17
13	0.33	-0.16	0.11	38	0.76	0.36	0.27
14	0.63	-0.24	0.24	39	0.84	0.53	0.33
15	0.71	0.42	0.20	40	0.42	0.68	0.11
16	0.32	0.53	0.11	41	0.47	1.30	0.22
17	0.47	1.15	0.10	42	0.56	-0.45	0.11
18	0.94	0.63	0.20	43	0.50	-0.24	0.11
19	0.66	1.25	0.15	44	0.82	0.43	0.19
20	0.61	-0.72	0.11	45	0.86	1.17	0.11
21	0.52	1.11	0.23	46	0.41	1.05	0.11
22	0.69	1.86	0.07	47	0.39	2.94	0.03
23	0.74	0.78	0.15	48	0.80	2.08	0.02
24	1.30	1.67	0.07	49	0.97	-0.05	0.49
25	0.32	1.58	0.11	50	0.41	0.04	0.11

Figure 1

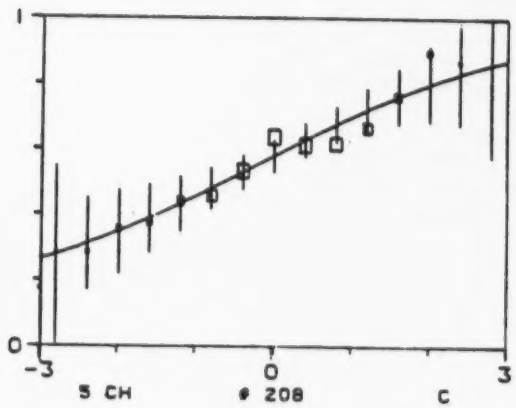
Examples of Item-Ability Regressions
Corresponding to Model Fit Scores of 0, 1, 2, and 3



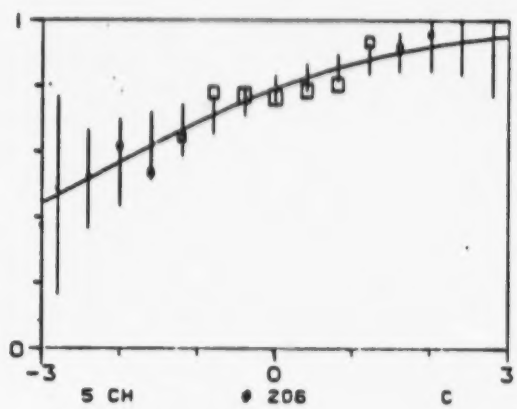
A



B



C



D

Figure 2

GMAT Verbal
Conversion Lines Based on Six Subpopulations

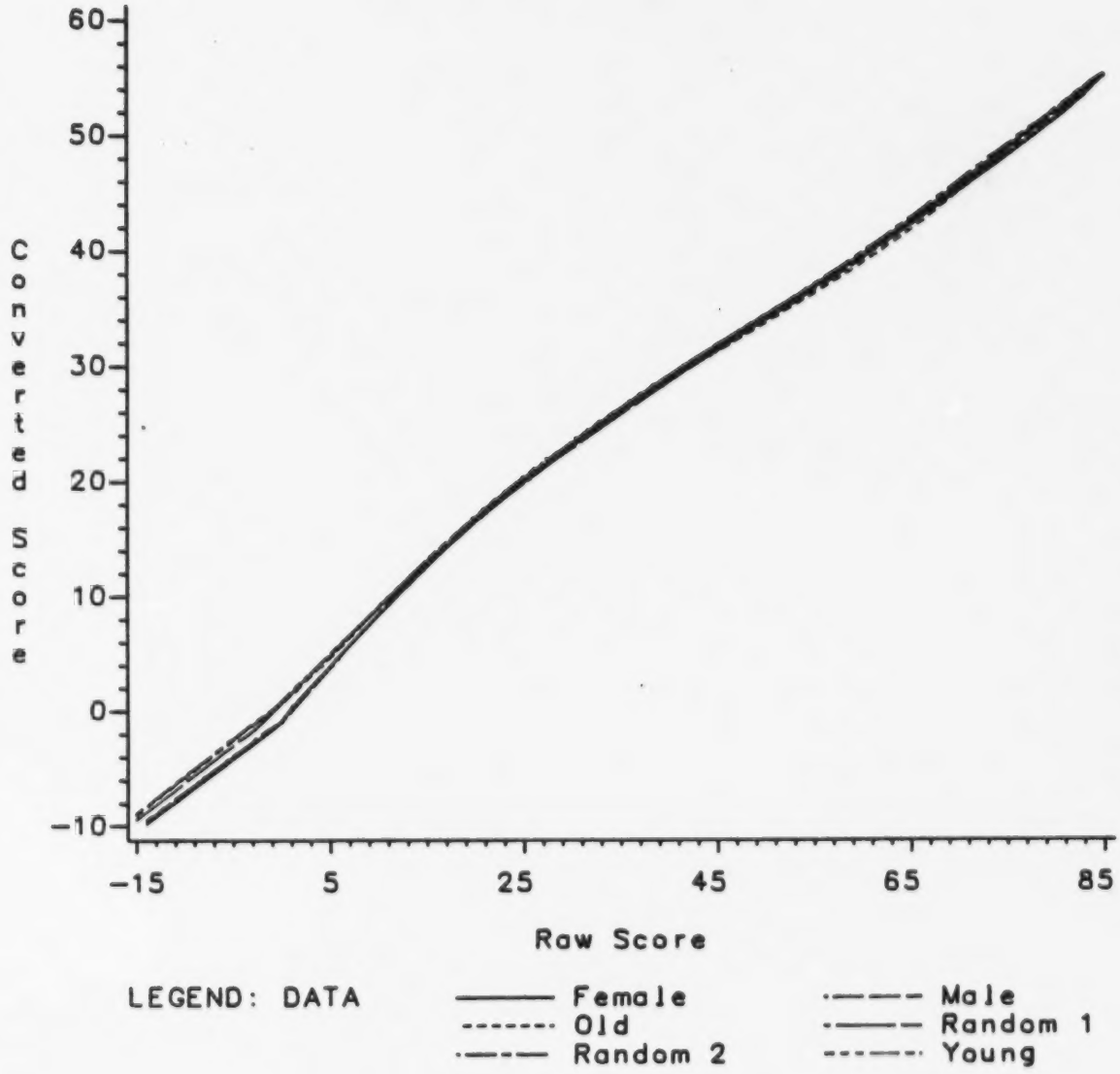


Figure 3

GMAT Verbal
Differences Between Conversion Lines

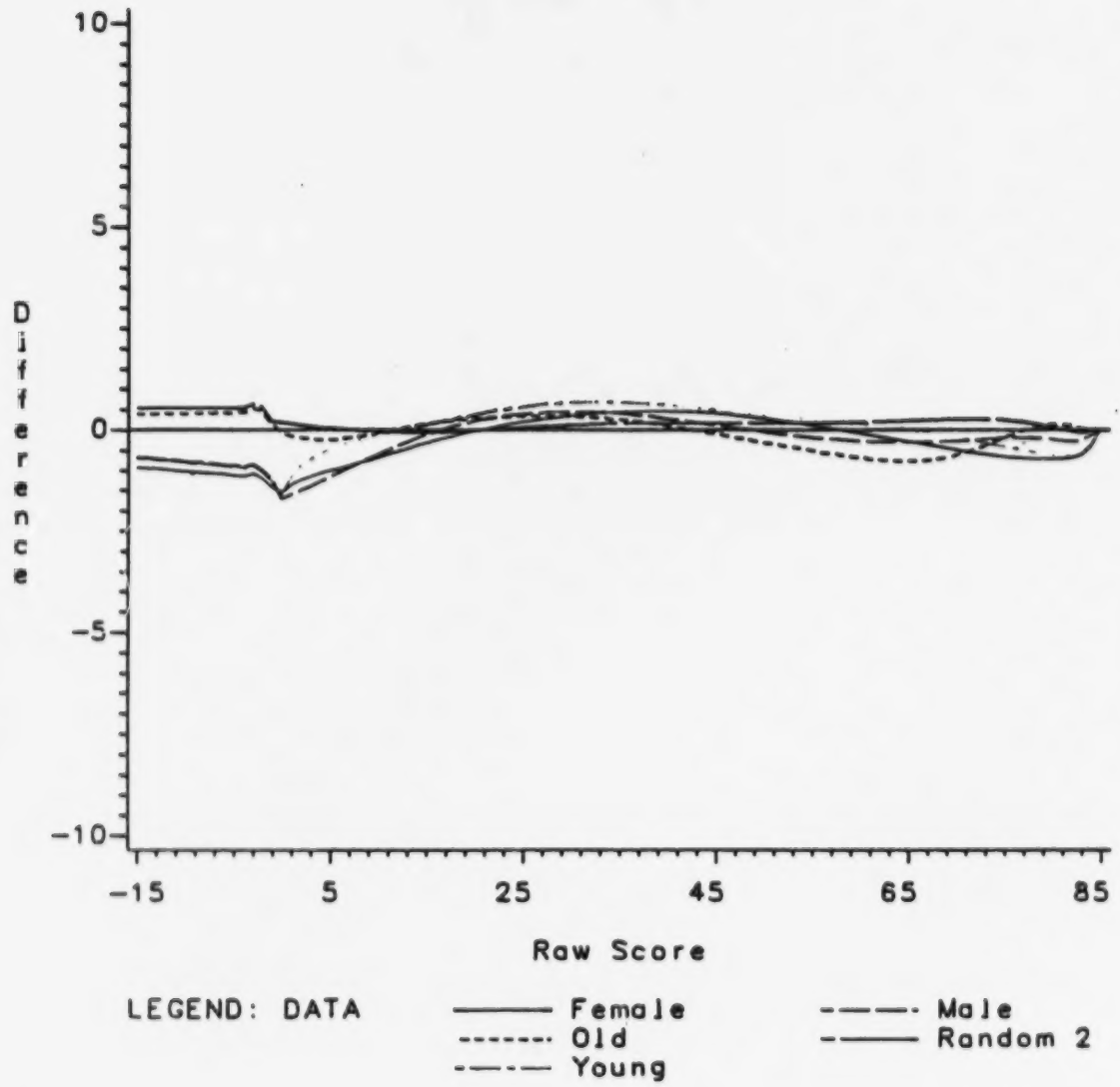


Figure 4

GMAT Quantitative
Conversion Lines Based on Six Subpopulations

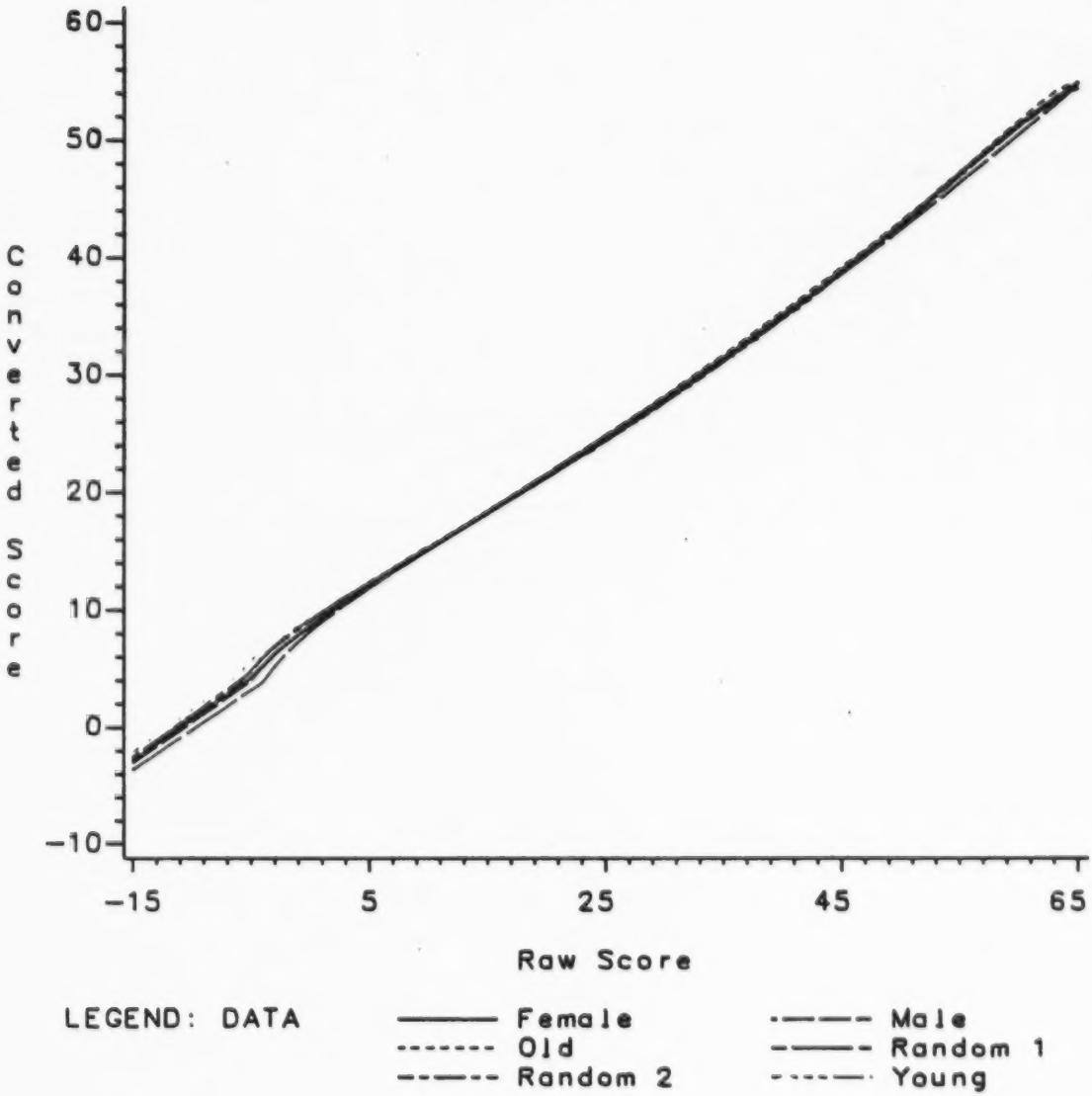


Figure 5

GMAT Quantitative
Differences Between Conversion Lines

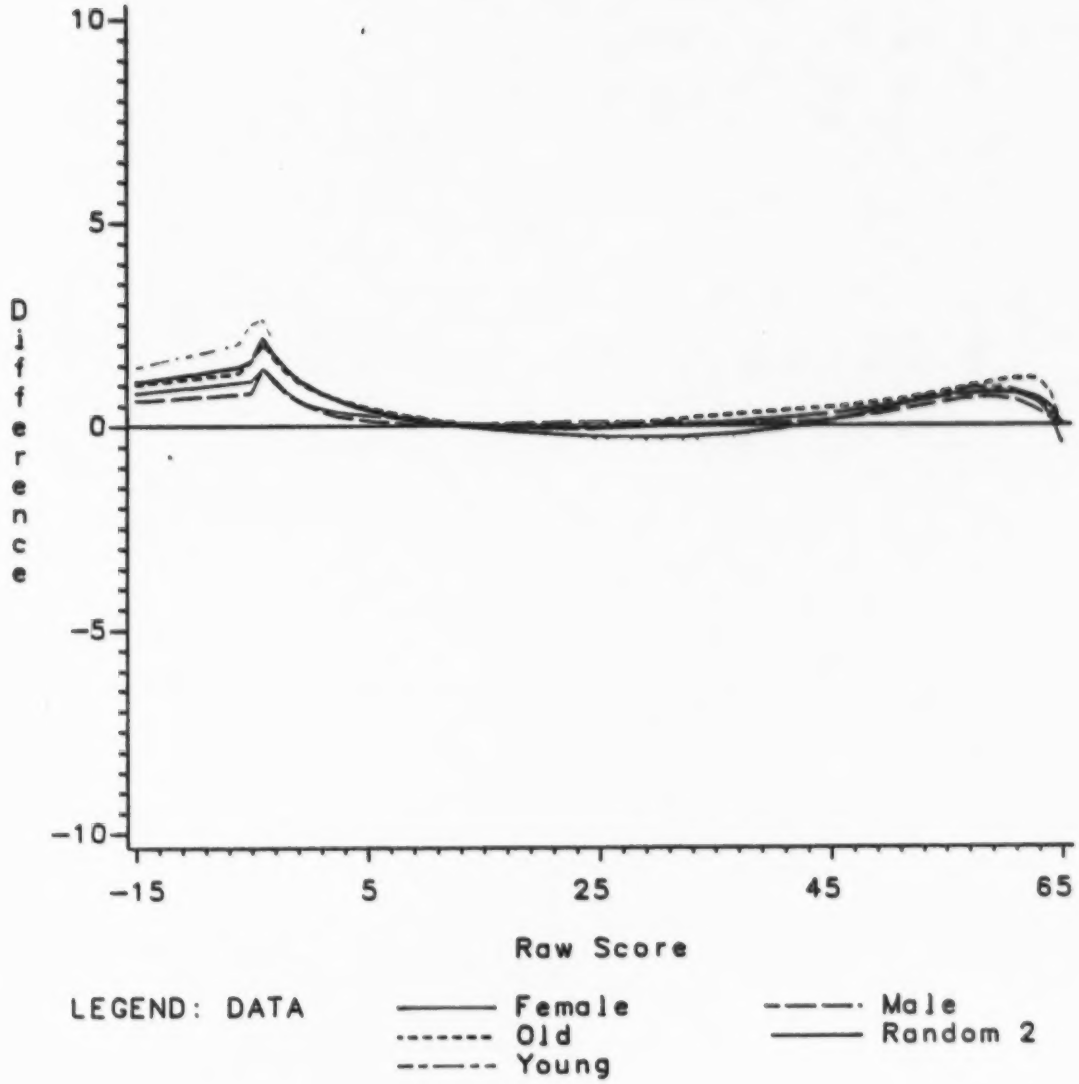


Figure 6

GMAT Verbal
Comparison of IRT Equating and SPE

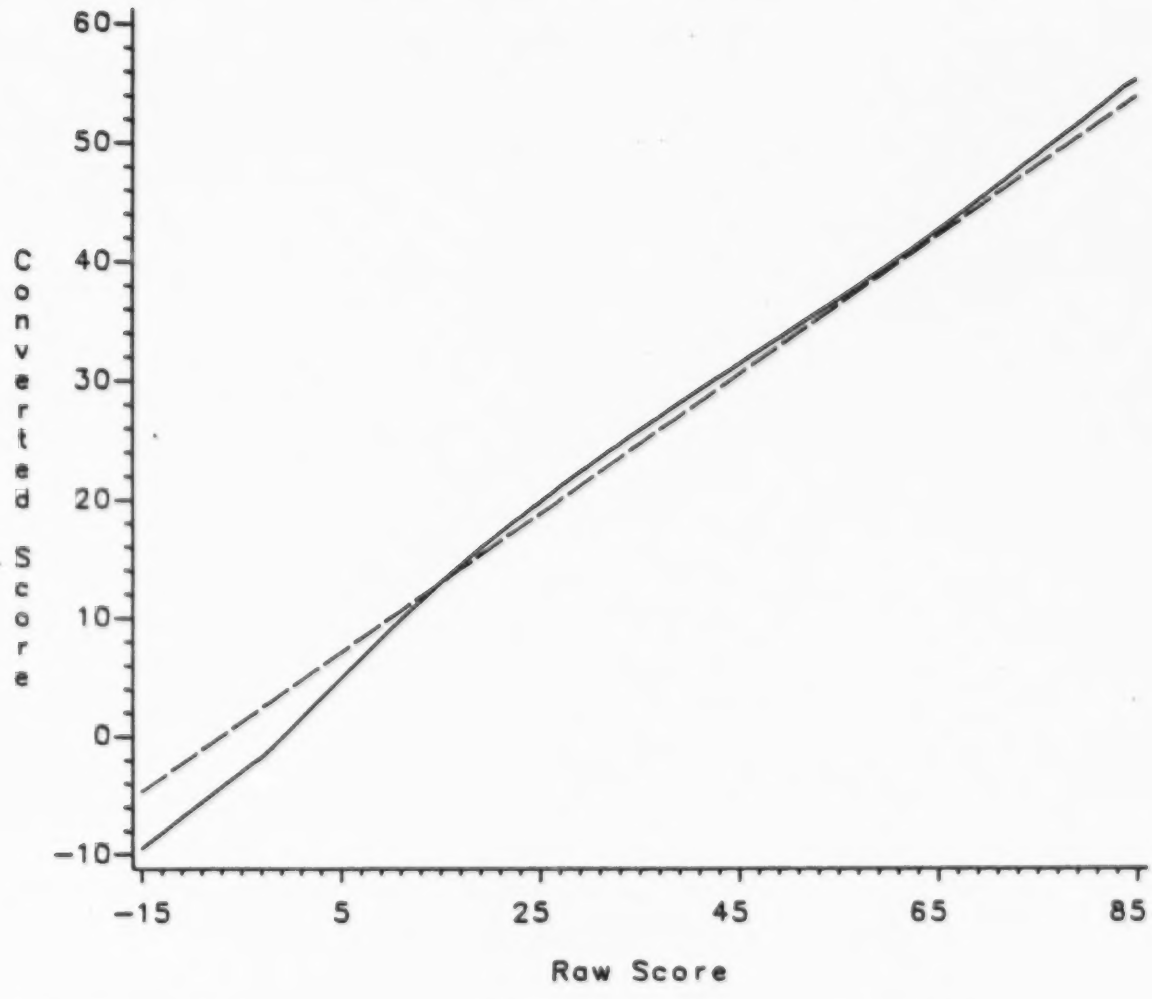


Figure 7

GMAT Quantitative
Comparison of IRT Equating and SPE